DOCUMENT RESUME

ED 385 582 TM 024 020

AUTHOR Kim, Sung-Ho
TITLE Instability in a Tree Approach to Regression. Program Statistics Research.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-92-1; ETS-TR-92-18
PUB DATE Jan 92
NOTE 39p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Error of Measurement; Models; *Regression (Statistics); Sample Size; *Selection; Simulation
IDENTIFIERS *Binary Trees

ABSTRACT

One of the major problems that a tree-approach to data analysis often encounters is the instability of tree-structures. The instability issue must be dealt with before data can be interpreted by this method. Examining instability at a node of a tree provides insight into the instability of the whole tree, because the same theory of instability applies to all the nodes. This paper deals with the instability issue at a single node of a tree. It is assumed that the data are from a regression model, and the factors in that model that affect the instability are examined. Squared-error loss is considered as a criterion for tree-construction (the "ls" criterion in the CART program). The selection rate of a regressor variable at a node of a tree is used as a measure of instability. The selection rate mainly depends on: (1) regression coefficients; (2) (conditional) variance-covariance structure of the regressor variables; (3) the sample size; and (4) noise in the response variable. Simulation results are reported that show patterns of instability for several different settings of regression models. Three figures and six tables illustrate the analysis. (Contains 10 references.) (SLD)

Instability In A Tree Approach To Regression

Sung-Ho Kim
Educational Testing Service

PROGRAM STATISTICS RESEARCH

TECHNICAL REPORT N 92-18

Educational Testing Service
Princeton, New Jersey 08541

Instability In A Tree Approach To Regression

Sung-Ho Kim
Educational Testing Service

Program Statistics Research
Technical Report No. 92-18

Research Report No. 92-1

Educational Testing Service
Princeton, New Jersey 08541

January 1992

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

INSTABILITY IN A TREE APPROACH TO REGRESSION

Sung-Ho Kim
Educational Testing Service
Princeton, NJ 08541

October 31, 1991

Thanks are due to Paul W. Holland for his advice and discussion with me that lead to this paper; Kikumi Tatsuoka and Howard Wainer for their comments until the final version of this paper. Elizabeth Brophy deserves my special thanks for her efficient typing.

Abstract

One of the major problems that a tree-approach to data analysis often encounters is *instability* of tree-structures. Thus if one wishes to interprete the data structure by the tree-approach, the instability issue must be dealt with.

Examining instability at a node of a tree provides insight into the instability of the whole tree, since the same theory of instability applies to all the nodes. Thus, this paper deals with the instability issue at a single node of a tree.

We assume that data are from a regression model, and examine what factors in that model affect the instability. Squared-error loss is considered as a criterion for tree-construction ("ls" criterion in CART program). The selection rate of a regressor variable at a node of a tree is used as a measure of instability. The selection rate mainly depends on (i) regression coefficients, (ii) (conditional) variance-covariance structure of the regressor variables (given a subset of the regressor variables), (iii) the sample size, and (iv) noise in the response variable. We report simulation results that show patterns of instability for several different settings of regression models.

1. INTRODUCTION AND MOTIVATION

In a typical sequential prediction procedure, we observe explanatory or predictor variables, one after another, deciding after each observation whether or not to continue adding variables. In selecting the next predictor variable, we usually attempt to maximize the expected utility, which involves the total cost of variable observations and the loss from the decision. This sequential procedure can be depicted by a directed acyclic graph, called a tree. We, however, refer to a tree-structured statistical prediction system as a tree. Variables are observed at the nodes of a tree.

Many of the presently available statistical techniques were designed for small data sets having standard structure with all variables of the same type; the underlying assumption was that the phenomenon is homogeneous. That is, that the same relationship between variables held over all of the measurement space. What makes a data set interesting is not only its size but also its complexity, where complexity can include such considerations as high dimensionality, a mixture of data types, nonstandard data structure and nonhomogeneity; that is, different relationships hold between variables in different parts of the measurement space. Tree-structured approaches have been suggested for data sets with such forms of complexity.

Use of trees in regression dates back to the AID (Automatic Interaction Detection) program developed by Morgan and Sonquist (1964). Then followed the ancestor classification program THAID, developed by Morgan and Messenger (1973). Breiman, Friedman, Olshen, and Stone (1984) proposed an algorithm called Classification and Regression Trees which is designed to provide a statistical sequential decision aid to its users for classification or regression problems. If we are given appropriate data, then we can get a guide, in a form of an upside-down tree, to what order to observe the predictor variables, when to stop observation, and what decision to make. The computer program that is based on this algorithm is referred to as CART. Huang (1989) developed a tree-structured method of detecting nonlinearity of a regression model. CART is now one of the most popular tree-structured data analysis and pattern recognition programs, and is used by many statisticians and AI people.

By the nature of the tree-structured approach, the approach is available for a data set which involves any large number of variables, where the variables can be of any type. It is also useful when the true regression model is non-linear, since it provides us a rough picture of the true model.

One of the advantages of the tree-structured approach is that the tree procedure output gives easily understood and interpreted information regarding the predictive structure of the data. The tree procedure output, almost universally, provides an illuminating and natural way of understanding the structure of the problem (Breiman et al. (1984), p. 58). However, extensive exploration and careful interpretation are necessary to arrive at sound conclusions (Einhorn (1972), Doyle (1973), Breiman et al. (1984)).

We will use the words "tree-shape" and "tree-structure" for different meanings. We define a tree-shape in terms of nodes and the directed arcs connecting the nodes. We define, for a given tree-shape, a tree-structure by assigning the selected predictor variable to each node and describing how to split the variable at the node. Figure 1.1 is an example of a tree-structure, where observations are made at the circles; decisions or predictions are made at the boxes. We use Figure 1.1 as follows. Suppose that the predictor variables are all binary, taking on the values 0 or 1. First, we observe the predictor variable X_1. If $X_1 = 1$, we stop observing and make prediction; otherwise, we observe X_2. The subsequent actions follow accordingly. If we delete all the letters and numbers from Figure 1.1, the remaining one is a tree-shape. We, however, use the terms "tree" and "tree-structure" in the same sense.

(Figure 1.1 about here)

Suppose we have a data set from a statistical model, and a tree is obtained based on the data set. With the sample size fixed, we repeat generating a data set from the same model and then obtaining a tree based on the data set. If the tree-structures are all the same over the repeated process, the tree-structures are said to be perfectly stable; otherwise, unstable with a level of instability, as will be discussed later in the paper.

We consider, for example, the tree in Figure 1.1. We label the node of X_i by the index i. Suppose there are several comparably informative variables at node 2. The variables appearing at node 3 will change according to the variables at node 2. A different variable

at node 2 may change the variables at the subsequent nodes. This phenomenon seems to erode the interpretability of the data structure by the tree approach. Is it really the case?

Breiman et al. discussed interpretability of the data structure via the tree output in their section 5.5. Instability of tree structures is a key issue there, and it certainly deserves a lot more investigation, since instability is a crucial obstacle to more sound interpretability. What are the factors that cause instability in trees? How do the factors affect instability? These issues will be investigated, in this paper, at a node of a tree under the assumption that the data are from a linear regression model. By seeing what factors involved in a regression model cause instability and how they do, we could have a better insight into the true statistical property behind data in the mist of instability. Understanding the instability issue at a node will give us an insight into the issue for a whole tree, since the same theory applies to all the nodes of a tree.

We consider a regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_r X_r + \epsilon \ , \tag{1.1}$$

where ϵ has $N(0, \sigma_\epsilon^2)$ distribution, and is independent of $(X_1, \cdots, X_r)$. We suppose we have a data set of size n from the model (1.1) such that the j^{th} observation is

$$(x_{j1}, \cdots, x_{jr}, y_j) \ .$$

For a vector or matrix A, A' means the transpose of A. We let

$$\begin{aligned} \underset{\sim}{Y}' &= (Y_1, \cdots, Y_n), \\ \underset{\sim}{X}' &= (\underset{\sim}{X}_1, \underset{\sim}{X}_2, \cdots, \underset{\sim}{X}_n) \ , \\ \underset{\sim}{\beta}' &= (\beta_0, \beta_1, \cdots, \beta_r), \\ \text{and} \quad \underset{\sim}{\epsilon}' &= (\epsilon_1, \epsilon_2, \cdots, \epsilon_n), \\ \text{where } \underset{\sim}{X}_j' &= (1, X_{j1}, X_{j2}, \cdots, X_{jr}), \quad \text{for } j = 1, 2, \cdots, n \ . \end{aligned} \tag{1.2}$$

Then, for a given data set (X, y), the ls estimate of β is given by

$$\hat{\beta} = (X'X)^{-1}X'y, \tag{1.3}$$

under the assumption that $X'X$ is of full rank. It is to be noted that $X_1, X_2, \cdots, X_r$ and Y are all assumed random variables.

The regression model as described above will be assumed throughout the paper. This paper consists of 5 sections. In Section 2, we introduce measures useful in dealing with the instability problem of the tree approach. The unbiased estimators of the measures introduced in Section 2 are derived in Section 3. In Section 4, instability of trees is illustrated using the unbiased estimators derived in Section 3. Finally, Section 5 presents several comments on the results of this paper.

2. MEASURES FOR THE TREE-STRUCTURED REGRESSION ANALYSIS.

Assume that all the X variables are finitely discrete or categorical. Suppose there is a data set generated from the model (1.1). Then, we have

$$V(Y) = \beta' \Sigma_X \beta + \sigma_\epsilon^2, \tag{2.1}$$

where $\beta' = (\beta_0, \beta_1, \beta_2, \cdots, \beta_r)$, and Σ_X is the variance-covariance matrix (VCM) of the column vector X, which is given by $X' = (1, X_1, X_2, \cdots, X_r)$.

Definition 2.1 Let $X_1, \cdots, X_r, Y$ be random variables.

For an integer s, $1 \le s \le r$, let $\{i_1, i_2, \cdots, i_s\}$ be a subset of $\{1, 2, \cdots, r\}$ and $X^* = (X_{i_1}, X_{i_2}, \cdots, X_{i_s})'$. Then, for $j \in \{1, 2, \cdots, r\} \backslash \{i_1, i_2, \cdots, i_s\}$, we let

$$\begin{aligned} IV_{X_j|X^*=x^*} &= V(Y|X^* = x^*) - E(V(Y|X^* = x^*, X_j)|X^* = x^*) \\ &= V(E(Y|X^* = x^*, X_j)|X^* = x^*). \end{aligned} \tag{2.2}$$

We call $IV_{X_j|\underset{\sim}{X}^* = \underset{\sim}{x}^*}$ the improvement value (IV) by X_j given $\underset{\sim}{X}^* = \underset{\sim}{x}^*$. If confusion is not likely, we will write $IV_{X_j|\underset{\sim}{x}^*}$ for $IV_{X_j|\underset{\sim}{X}^* = \underset{\sim}{x}^*}$. In the tree approach, we analyze the relationship between Y and the set of the X-variables by selecting the X-variables one after another. At the initial selection, select the X-variable for which

$$IV_X = V(E(Y|X)) \tag{2.3}$$

is maximized. Let the selected variable be X_1. Then, for $X_1 = x_1$, say, repeat the same process. That is, select the X-variable for which

$$IV_{X|X_1 = x_1} = V(E(Y|X, X_1 = x_1)|X_1 = x_1) \tag{2.4}$$

is maximized. If such difference as in (2.4), say, is equal to zero, then we stop the selection process.

A careful look at the IV would give us an insight into the relationship between the tree-structure and the regression model. At this point, we need the theorem below. For notational convenience, we will use X_0 for the first element ($=1$) of $\underset{\sim}{X}$.

Theorem 2.2

Suppose the following two conditions hold for the regression model (1.1):

(i) $\underset{\sim}{X} = (X_0, X_1, \cdots, X_r)'$ is a random vector with a VCM $\Sigma_{\underset{\sim}{X}}$,

(ii) the coefficients $\beta_0, \beta_1, \cdots, \beta_r$ are known.

Then, under the set-up of Definition 2.1, we have

$$IV_{X_j|\underset{\sim}{x}^*} = \underset{\sim}{\beta}' \Sigma_{\underset{\sim}{X}|\underset{\sim}{x}^*} \underset{\sim}{\beta} - \underset{\sim}{\beta}' E(\Sigma_{\underset{\sim}{X}|\underset{\sim}{x}^*, X_j}) \underset{\sim}{\beta} \tag{2.5}$$

where $\Sigma_{\underset{\sim}{X}|\underset{\sim}{X}^* = \underset{\sim}{x}^*}$ is the VCM of $\underset{\sim}{X}$ conditional on that $\underset{\sim}{X}^* = \underset{\sim}{x}^*$.

Proof: Its proof is straightforward from the regression model (1.1).

$$V(Y|\underset{\sim}{X}^* = \underset{\sim}{x}^*) = V(\underset{\sim}{X}'\underset{\sim}{\beta} + \epsilon|\underset{\sim}{X}^* = \underset{\sim}{x}^*)$$
$$= V(\underset{\sim}{X}'\underset{\sim}{\beta}|\underset{\sim}{X}^* = \underset{\sim}{x}^*) + \sigma_\epsilon^2$$
$$= \underset{\sim}{\beta}'\,\underset{\sim}{\Sigma}_{\underset{\sim}{X}|\underset{\sim}{X}^* = \underset{\sim}{x}^*}\,\underset{\sim}{\beta} + \sigma_\epsilon^2 .$$

Similarly, we have

$$V(Y|\underset{\sim}{X}^* = \underset{\sim}{x}^*, X_j = x_j) = \underset{\sim}{\beta}'\,\underset{\sim}{\Sigma}_{\underset{\sim}{X}|\underset{\sim}{X}^* = \underset{\sim}{x}^*, X_j = x_j}\,\underset{\sim}{\beta} + \sigma_\epsilon^2 .$$

By Definition 2.1, (2.5) follows.□

If confusion is not likely, we will write $\underset{\sim}{\Sigma}_{\underset{\sim}{X}|\underset{\sim}{x}^*}$ for $\underset{\sim}{\Sigma}_{\underset{\sim}{X}|\underset{\sim}{X}^* = \underset{\sim}{x}^*}$. Theorem 2.2 says that the *IV* depends upon the regression coefficients and the (conditional) *VCM* of $\underset{\sim}{X}$ (given a subset of $\{X_1, \cdots, X_r\}$).

IV_{X_i}, for the initial selection of X_i variable, is given by

$$IV_{X_i} = \underset{\sim}{\beta}'\,\underset{\sim}{\Sigma}_{\underset{\sim}{X}}\,\underset{\sim}{\beta} - \underset{\sim}{\beta}'\,E(\underset{\sim}{\Sigma}_{\underset{\sim}{X}|X_i})\,\underset{\sim}{\beta} . \tag{2.6}$$

The variation among the *IV*'s deserves our attention since it has something to do with instability of trees. The following corollary is immediate from Theorem 2.2, and thus proof-omitted.

Corollary 2.3

Under the same set-up of Theorem 2.2, for j and j' $(j \neq j)$, both in $\{1, 2, \cdots, r\} / \{i_1, i_2, \cdots, i_s\}$, we have

$$IV_{X_j|\underset{\sim}{x}^*} - IV_{X_{j'}|\underset{\sim}{x}^*} = \underset{\sim}{\beta}'\,E(\underset{\sim}{\Sigma}_{\underset{\sim}{X}|\underset{\sim}{x}^*, X_{j'}})\,\underset{\sim}{\beta} - \underset{\sim}{\beta}'\,E(\underset{\sim}{\Sigma}_{\underset{\sim}{X}|\underset{\sim}{x}^*, X_j})\,\underset{\sim}{\beta} . \tag{2.7}$$

In particular, if $r = 2$ in the regression model (1.1), we have, from (2.7), that

$$IV_{X_1} - IV_{X_2} = \beta_1^2 \, E(V(X_1|X_2)) - \beta_2^2 \, E(V(X_2|X_1)) \, . \tag{2.8}$$

From equation (2.8), we can at least say that, the larger β_1^2 or $E(V(X_1|X_2))$, the higher would the probability be that X_1 is selected rather than X_2.

In this section, we have found an expression for $IV_{X|\cdot}$ under the condition of Theorem 2.2.

3. UNBIASED ESTIMATORS

In this section, I will derive an unbiased estimator of the $IV_{X|\cdot}$ introduced in Section 2.

Lemma 3.1 Let W_1, ⁻, W_n be random variables, and A an $n \times n$ matrix. If $\underset{\sim}{W} = (W_1, \cdots, W_n)'$, then

$$E(\underset{\sim}{W}' A \underset{\sim}{W}) = E(\underset{\sim}{W}') A \, E(\underset{\sim}{W}) + \sum_{i,j} a_{ij} \, cov(W_i, W_j),$$

where a_{ij} is the $(i, j)^{th}$ entry of A.

Proof: From the equation

$$\underset{\sim}{W}' A \underset{\sim}{W} = \sum_{i,j} W_i \, W_j \, a_{ij},$$

the desired result is a straightforward consequence.□

Theorem 3.2

Let B be a $(r + 1) \times (r + 1)$ matrix. Then, given the data $(X, \underset{\sim}{y})$ from the regression model (1.1),

$$E(\hat{\beta}_X \, B \, \hat{\beta}_X | X) = \beta \, B \, \beta + \sigma_\epsilon^2 \, \mathrm{tr}(B(X'X)^{-1}). \tag{3.1}$$

Proof: By substituting (1.3) in (3.1), we have

$$\begin{aligned} E(\hat{\beta}_X \, B \, \hat{\beta}_X | X) &= E(Y \, X \, (X'X)^{-1} \, B \, (X'X)^{-1} X' Y | X) \\ &= \beta \, X'X(X'X)^{-1} \, B(X'X)^{-1} X'X\beta + c \\ &= \beta \, B \, \beta + c, \end{aligned}$$

where

$$\begin{aligned} c &= \sigma_\epsilon^2 \, \mathrm{tr}(X(X'X)^{-1} \, B(X'X)^{-1} \, X') \\ &= \sigma_\epsilon^2 \, \mathrm{tr}(B(X'X)^{-1}), \end{aligned}$$

by Lemma 3.1. Q.E.D.

Suppose we have a data set of size n from the model (1.1). Let I_n be the $n \times n$ identity matrix, and J_n the $n \times n$ matrix of 1's. Then we have

$$\frac{1}{n-1} \, \mathrm{E}\left[X'(I_n - \frac{1}{n} J_n)X\right] = \Sigma_X . \tag{3.2}$$

For the given data set, suppose that n_{j,x_j} cases have $X_j = x_j$, then the summation of n_{j,x_j} over all the possible values of x_j of X_j is equal to n. Let $X_{(j,x_j)}$ be the $n_{j,x_j} \times (r+1)$ matrix composed of the rows of X each of whose $(j+1)^{\text{th}}$ entries is x_j.

In analogy to (3.2), we have

$$\frac{1}{n_{j,x_j} - 1} \, \mathrm{E}(X'_{(j,x_j)} \, (I_{n_{j,x_j}} - \frac{1}{n_{j,x_j}} J_{n_{j,x_j}}) \, X_{(j,x_j)}) = \Sigma_{X|X_j = x_j} .$$

We define

$$\hat{\Sigma}_X = \frac{1}{n-1} X'(I_n - \frac{1}{n} J_n)X \ , \tag{3.3}$$

and

$$\hat{\Sigma}_{X|X_j = x_j} = \frac{1}{n_{j,x_j} - 1} X'_{(j,x_j)} (I_{n_{j,x_j}} - \frac{1}{n_{j,x_j}} J_{n_{j,x_j}})X_{(j,x_j)} \ . \tag{3.4}$$

Then, we can see that, if X_j and the other X-variable are not independent,

$$\frac{1}{n} \sum_{x_j} n_{j,x_j} \hat{\Sigma}_{X|X_j = x_j} = \hat{\Sigma}_{X|X_j} \ , \text{ say,} \tag{3.5}$$

is an unbiased estimator of $E(\Sigma_{X|X_j})$; otherwise $\hat{\Sigma}_{X|X_j}$ is given by (3.8) below.

In (3.5), the summation is done over the support set of X_j.

We let $\underset{\sim}{i} = (i_1, i_2, \cdots, i_s)$, for $1 \le s \le r$, and let $D^{(i)}$ be the $(r + 1) \times (r + 1)$ diagonal matrix where

$$\text{the } (j, j)^{\text{th}} \text{ entry} = \begin{cases} 0 & \text{if } j \in \{ i_1, \cdots, i_s\} \\ 1 & \text{otherwise .} \end{cases}$$

We let $\alpha = \{1, 2, \cdots, r\} \backslash \{i_1, i_2, \cdots, i_s\}$. If X^* and X_j are independent for $j \in \alpha$, then

$\Sigma_{X|x^*} = D^{(i)} \Sigma_X D^{(i)}$ for each possible value x^* of X^*.

Recall that the matrix X in expression (1.2) is a random matrix. For a given set of data (X, y), suppose we fit the linear regression model (1.1), and the ls estimator of β is denoted by $\hat{\beta}_X$.

Under the normality assumption of $\underset{\sim}{\epsilon}$ and the independence assumption of $X_1, \cdots, X_r$, the mean squared error (MSE) from the least-square fit of the model (1.1) is the unique minimum variance unbiased estimator of σ_ϵ^2 (Atiqullah (1962)). We will denote the MSE by $\hat{\sigma}_\epsilon^2$. We can also find the uniformly minimum varianced unbiased (UMVU) estimator of σ_ϵ^2, when $X_1, \cdots, X_r$ are correlated (Theorem 4.1 of Lehmann (1983)). We also denote the estimator by $\hat{\sigma}_\epsilon^2$.

The *IV* value in Definition 2.1 depends on the joint distribution of $X_1, X_2, \cdots, X_r$ and Y. If we base the *IV* value on the X-matrix of a given data $(X, \underset{\sim}{y})$, and denote such *IV* by IV^X, then we may write

$$IV^X_{X_j|x^\bullet} = \underset{\sim}{\beta}' \underset{\sim}{\Sigma}_{X|x^\bullet} \underset{\sim}{\beta} - \underset{\sim}{\beta}' \underset{\sim}{\Sigma}_{X|x^\bullet, X_j} \underset{\sim}{\beta} \tag{3.6}$$

Theorem 3.3

Suppose the regression model (1.1) is true. Then, given the data $(X, \underset{\sim}{y})$, the statistic given below is an unbiased estimator of $IV^X_{X_j|x^\bullet}$, for $j \in \alpha$:

$$IV^X_{X_j|x^\bullet} = \underset{\sim}{\hat{\beta}}'_X \left(\underset{\sim}{\Sigma}_{X|x^\bullet} - \underset{\sim}{\Sigma}_{X|x^\bullet, X_j}\right) \underset{\sim}{\hat{\beta}}_X - \hat{\sigma}_\epsilon^2 \operatorname{tr}\left((\underset{\sim}{\Sigma}_{X|x^\bullet} - \underset{\sim}{\Sigma}_{X|x^\bullet, X_j})(X'X)^{-1}\right). \tag{3.7}$$

Proof: Under the normality assumption of $\underset{\sim}{\epsilon}$, we can always find the UMVU estimator $\hat{\sigma}_\epsilon^2$ of σ_ϵ^2. The rest of the theorem follows immediately from Theorem 3.2. □

We suppose that

[IND-1] for $j \in \alpha$, $\underset{\sim}{X}^{*}$, X_j, and the vector of the rest of the X-variables are mutually independent.

Then,

$$\underset{\sim}{\Sigma}_{X|x^{*}} = \underset{\sim}{D}^{(i)} \underset{\sim}{\Sigma}_{X} \underset{\sim}{D}^{(i)} \text{ and } \underset{\sim}{\Sigma}_{X|x^{*}, X_j} = \underset{\sim}{D}^{(i,j)} \underset{\sim}{\Sigma}_{X} \underset{\sim}{D}^{(i,j)} . \tag{3.8}$$

Thus in the mutual independence situation, we have, from (2.5),

$$IV_{X_j|x^{*}} = \underset{\sim}{\beta}' \underset{\sim}{D}^{(i)} \underset{\sim}{\Sigma}_{X} \underset{\sim}{D}^{(i)} \underset{\sim}{\beta} - \underset{\sim}{\beta}' \underset{\sim}{D}^{(i,j)} \underset{\sim}{\Sigma}_{X} \underset{\sim}{D}^{(i,j)} \underset{\sim}{\beta} .$$

$IV^{X}_{X_j|x^{*}}$ thus be written as follows:

$$IV^{X}_{X_j|x^{*}} = \underset{\sim}{\beta}' \underset{\sim}{D}^{(i)} \underset{\sim}{\hat{\Sigma}}_{X} \underset{\sim}{D}^{(i)} \underset{\sim}{\beta} - \underset{\sim}{\beta}' \underset{\sim}{D}^{(i,j)} \underset{\sim}{\hat{\Sigma}}_{X} \underset{\sim}{D}^{(i,j)} \underset{\sim}{\beta}. \tag{3.9}$$

Consequently, we have the following result.

Theorem 3.4

Suppose the regression model (1.1) is true. Then, under the independence condition [IND-1] and given the data $(X, \underset{\sim}{y})$ from the model (1.1), the statistic given below is an unbiased estimator of $IV^{X}_{X_j|x^{*}}$ in expression (3.9):

$$\widehat{IV}^{X}_{X_j|x^{*}} = \underset{\sim}{\hat{\beta}}'_{X} \left(\underset{\sim}{D}^{(i)} \underset{\sim}{\hat{\Sigma}}_{X} \underset{\sim}{D}^{(i)} - \underset{\sim}{D}^{(i,j)} \underset{\sim}{\hat{\Sigma}}_{X} \underset{\sim}{D}^{(i,j)} \right) \underset{\sim}{\hat{\beta}}_{X} - K, \tag{3.10}$$

where

$$K = \hat{\sigma}^2_{\epsilon} \left(\sum_{l \in \alpha} S_{jl} (X'X)^{-1}_{lj} + \sum_{k \in \alpha_j} S_{kj} (X'X)^{-1}_{jk} \right), \tag{3.11}$$

S_{kl} is the sample covariance of X_k and X_l,

A_{kl} is the $(k, l)^{\text{th}}$ element of matrix A, and $\alpha_j = \alpha \backslash \{j\}$.

Proof: The proof is sufficient if we show equation (3.11).

$$\begin{aligned} & \mathrm{tr}(D^{(i)} \hat{\Sigma}_X D^{(i)} (X'X)^{-1}) \\ &= \frac{1}{n-1} \mathrm{tr}\left(D^{(i)} X' (I_n - \frac{1}{n} J_n) X D^{(i)} (X'X)^{-1} \right) \qquad \text{by (3.2)} \\ &= \frac{1}{n-1} \mathrm{tr}\left(D^{(i)} X'X D^{(i)} (X'X)^{-1} \right) - \frac{1}{n(n-1)} \mathrm{tr}\left(D^{(i)} X' J_n X D^{(i)} (X'X)^{-1} \right) . \end{aligned} \tag{3.12}$$

For the first term in (3.12);

$$\begin{aligned} \mathrm{tr}\left(D^{(i)} X'X D^{(i)} (X'X)^{-1} \right) &= \mathrm{tr}\left((D^{(i)} X'X D^{(i)}) (D^{(i)} (X'X)^{-1} D^{(i)}) \right) \\ &= \sum_{k \in \alpha} \sum_{l \in \alpha} (X'X)_{kl} (X'X)^{-1}_{lk} . \end{aligned} \tag{3.13}$$

For the second term in (3.12);

$$\begin{aligned} \mathrm{tr}\left(D^{(i)} X' J_n X D^{(i)} (X'X)^{-1} \right) &= \mathrm{tr}\left((D^{(i)} X' J_n X D^{(i)}) (D^{(i)} (X'X)^{-1} D^{(i)}) \right) \\ &= \sum_{k \in \alpha} \sum_{l \in \alpha} (X' J_n X)_{kl} (X'X)^{-1}_{lk} . \end{aligned} \tag{3.14}$$

After a simple algebra, we have

$$X' J_n X = n^2 M, \tag{3.15}$$

where M is the $(r + 1) \times (r + 1)$ matrix, with its $(i + 1, j + 1)^{\text{th}}$ entry being

$$\left(\sum_{k=1}^{n} X_{ki} \right) \left(\sum_{k=1}^{n} X_{kj} \right) / n^2 .$$

From (3.14) and (3.15), we have

$$\mathrm{tr}\left(D^{(i)}_{\sim} X' J_n X D^{(i)}_{\sim} (XX)^{-1}\right) = n^2 \sum_{k\in\alpha} \sum_{l\in\alpha} M_{kl} (X'X)^{-1}_{lk} . \tag{3.16}$$

By (3.12), (3.13), and (3.16), we have

$$\begin{aligned}(n-1)\ \mathrm{tr}\left(D^{(i}_{\sim} \hat{\Sigma}_{X} D^{(i)}_{\sim} (X'X)^{-1}\right) &= \sum_{k\in\alpha} \sum_{l\in\alpha} \left((X'X)_{kl} - nM_{kl}\right)(X'X)^{-1}_{lk} \\ &= (n-1) \sum_{k\in\alpha} \sum_{l\in\alpha} S_{kl}\, (X'X)^{-1}_{lk} .\end{aligned} \tag{3.17}$$

By the same argument, we have

$$(n-1)\ \mathrm{tr}\left(D^{(i,j)}_{\sim} \hat{\Sigma}_{X} D^{(i,j)}_{\sim} (X'X)^{-1}\right) = (n-1) \sum_{k\in\alpha_j} \sum_{l\in\alpha_j} S_{kl}\, (X'X)^{-1}_{lk} . \tag{3.18}$$

From (3.17) and (3.18) follows

$$\begin{aligned}&\mathrm{tr}\left(D^{(i)}_{\sim} \hat{\Sigma}_{X} D^{(i)}_{\sim} (X'X)^{-1}\right) - \mathrm{tr}\left(D^{(i,j)}_{\sim} \hat{\Sigma}_{X} D^{(i,j)}_{\sim} (X'X)^{-1}\right) \\ &\quad = \sum_{l\in\alpha} S_{jl}\, (X'X)^{-1}_{lj} + \sum_{k\in\alpha_j} S_{kj}\, (X'X)^{-1}_{jk} .\end{aligned}$$

Therefore, by Theorem 3.3 and expression (3.8), we get the desired result. □

It is noteworthy that the unbiased estimator of $IV^{X}_{X_j|x^{\bullet}_{\sim}}$ in Theorem 3.3 depends on the ls estimator of $\underset{\sim}{\beta}$ based on the whole data $(X, \underset{\sim}{y})$ rather than based on any subset of the data $(X, \underset{\sim}{y})$ corresponding to the outcome $X^{\bullet} = x^{\bullet}$. Meanwhile, the estimator depends on the conditional covariance structure of the X-variables.

Under the independence assumption of $X_1, X_2, \cdots, X_r$ the bias term K is non-negative for large n. Thus ignoring the bias term results in overestimation. Actually, we may take $S_{kl} = 0$, $k \neq l$, for large n. Then, from (3.11),

$$K = \hat{\sigma}_\epsilon^2 \, S_{jj} \, (X'X)_{jj}^{-1} \geq 0 \, ,$$

since the independence assumption implies that $(X'X)^{-1}$ is positive definite, and so the diagonal elements are all positive.

4. ILLUSTRATIONS

In this section, we will see some simple examples of instability for several causal factors of it which are discussed in Section 3. Tree-structures may be unstable partially due to chance fluctuations in the data or due to associations between the variables (see Subsections 5.5.2 and 8.10.1 of Breiman, et al. (1984)). The last paragraph of Subsection 8.10.1 may have to be read with discretion. For the regression model used in their Section 8.6, the regression coefficients are all different by some amount, while the variances of the X-variables are all within a small range. In this situation, the tree-structure may be very stable, as will be shown in Example 4.1.

Example 4.1

Consider a regression model (1.1) with $r = 2$, and suppose that the X-variables are independent. If there are no X-variables already known, i.e., $\{i_1, i_2, \cdots, i_s\} = \phi$, then, from (3.10), we have

$$\widehat{IV}_{X_j}^{X} = \hat{\beta}_X' \, (\hat{\Sigma}_X - D^{(j)} \, \hat{\Sigma}_X \, D^{(j)}) \hat{\beta}_X - \hat{\sigma}_\epsilon^2 \, S_{jj} \, (X'X)_{jj}^{-1} \, . \tag{4.1}$$

Since there are only two X-variables, we may look at

$$\mathrm{DIV}_{1,2} = \widehat{IV}_{X_1}^{X} - \widehat{IV}_{X_2}^{X}$$

to see which X-variable is actually selected based on a given data set. From (4.1) follows

$$\mathrm{DIV}_{1,2} = \hat{\beta}_1^2 \, S_{11} - \hat{\beta}_2^2 \, S_{22} + \hat{\sigma}_\epsilon^2 \left(S_{22}(X'X)_{33}^{-1} - S_{11}(X'X)_{22}^{-1} \right). \tag{4.2}$$

For simulation, we consider a version of the regression model (1.1) (call it M-1) under the following conditions:

(a) $\beta_0 = \beta_1 = \beta_2 = 1$,

(b) $\sigma_\epsilon^2 = 1$,

(c) $P(X_1 = 1) = 0.2$, $P(X_1 = 2) = 0.8$, $P(X_2 = 1) = P(X_2 = 2) = 0.5$.

If the DIV-value is positive, then we select X_1 variable; if negative, X_2 being selected. If the DIV is equal to zero, then both variables are equally likely. This selection rule is the same as the CART's with the "least-square" selection criterion of CART. Table 4.1 is obtained based on 10 data sets of size 100 each from the model M-1. Each row corresponds to each data set.

(Table 4.1 about here)

As indicated in Table 4.1, there is some uncertainty in variable-selection. To get some idea of uncertainty, we generated 500 data sets of size 30 each. Figure 4.1 is the histogram of the 500 *DIV*-values.

(Figure 4.1 about here)

Table 4.2 shows the selection rates of X_1 variable out of 1,000 iterations for each specified regression model. For the table, we allowed 1, 2, and 3 for β_1; (0.2, 0.3), (0.2, 0.5), (0.3, 0.4), and (0.3, 0.5) for (P(X_1 = 1), P(X_2 = 1)); 5, 10, 30, and 50 for the sample size. The values in the row of E(DIV) (call it the "true DIV") are obtained from $IV_{X_1} - IV_{X_2}$.

(Table 4.2 about here)

(Figure 4.2 about here)

Table 4.2 is graphed in Figure 4.2, where the numbers on the right margin or on the lines are the true DIV values. From the graph we can see that the selection rate depends on the true DIV and the sample size. When the true DIV is larger than or equal to 0.39, the selection rate is not less than 0.75 even at the sample size 10. On the other hand, for the true DIV's between -0.09 and -0.03, the selection rate is not less than 0.3 even for the sample size 50.

(Table 4.3 about here)

Table 4.3 shows the relationship between the selection rate of X_1 and the sample size for the regression model M-1. According to the table, we need a sample of size larger than 300 to reach the selection rate of X_1 0.1, and 600 to reach the selection rate 0.05. This is an extreme situation compared with the case where $\beta_1 = 2$ or 3 in Table 4.2.□

We may safely conclude from Example 4.1 that the absolute distance between the selection rate and 0.5 increases

(i) as the absolute value of the true DIV increases for each sample size, or

(ii) as the sample size increases for each true DIV.

We define the level of instability or the instability level (at a node) to be equal to 0.5 minus the above-mentioned absolute distance. Thus the instability level is between 0 and 0.5 inclusive. O means "the lowest instability (i.e., perfect stability)", and 0.5 "the highest instability".

As implied by expression (3.7), the level of instability depends on the sample size, the association level among the X-variables, σ_ϵ^2, and $\underset{\sim}{\beta}$. If we knew $IV_{X_j|\underset{\sim}{x}}$ for any subset $\{X_{i_1}, \cdots, X_{i_s}\}$ of $\{X_1, \cdots, X_r\}$ and $j \in \alpha = \{1, 2, \cdots, r\}\backslash\{i_1, i_2, \cdots, i_s\}$, then from the tree Υ_∞ which is obtained based on $IV_{X_j|\underset{\sim}{x}}$ could we see which X-variable partitions the population so that the partitioned subgroups are most homogeneous with respect to Y, i.e., the within-group variances of Y are minimized; and so on, for all the subsequent nodes. That is, conditional on that a set of X-variables are already observed at the previous nodes, we select the X-variable which divides the current subset of the population into mostly homogeneous subgroups. If we say that Υ_∞ is an unknown parameter, then we may say that the tree $\Upsilon_{\underset{\sim}{X}, \underset{\sim}{y}}$ which is obtained based on the data $(\underset{\sim}{X}, \underset{\sim}{y})$ is an estimate of the parameter. As indicated in Example 4.1, we can expect that the tree $\Upsilon_{\underset{\sim}{X}, \underset{\sim}{y}}$ will approach Υ_∞ as the sample size increases.

However, Υ_∞ may not be an interesting object, because Υ_∞ does not necessarily show the whole picture of the corresponding statistical model. This is analogous to that the scatterplots of all the pairs of Y and X variables do not reveal the joint structure of the data. In some sense, instability of trees can be a signal to data analysts that further investigation is desirable on data.

The example below is continued from Example 4.1, and illustrates how the noise in the regression model affects instability of trees.

Example 4.2

Consider a regression model (1.1) with $r = 2$, which satisfies condition (c) for model (M-1) of Example 4.1, and $\beta_0 = \beta_2 = 1$. In this example, we allow 1, 2, and 3 for β_1, and 2, 3, and 4 for σ_ϵ, and see how the selection rate of X_1 changes. The selection rates for $\sigma_\epsilon = 1$ are in Table 4.2. Table 4.4 is obtained by the same method as for Table 4.2 (the number of repeat = 1,000). From (3.6), we can see that the true DIV has nothing to do with the noise (σ_ϵ). Expression (4.2) says that the noise affects the tree-instability through the bias term.

Table 4.4 says that instability of trees becomes serious as the noise (σ_ϵ) to the response variable increases. From Table 4.4 and the fourth column of Table 4.2, we can see that the selection rate of X_1 gets closer to 0.5 as the noise increases for each sample size. Expression (3.7) explains this phenomenon. But, since $(X'X)^{-1}$ converges in the order of $0\left(\frac{1}{n}\right)$, the instability due to the noise (σ_ϵ) can be overcome by increasing the sample size only.

(Table 4.4 about here)

□

Next, we will consider a case where X-variables are associated.

Example 4.3

Consider two versions of the regression model (1.1) with r=2, where the X's are binary (0 or 1), and the two models differ in the joint probability of the X's. Their joint probabilities are given by Table 4.5 (a) and (b), respectively. We call the model corresponding to Table 4.5 (a) by model (M-2a); Table 4.5 (b) by model (M-2b).

(Table 4.5 about here)

We put

(a) $\beta_0 = \beta_2 = 1$, and

(b) $\sigma_\epsilon^2 = 1$.

The marginals of X_1 and X_2 for both models are as in (c) of model (M-1) of Example 4.1. We allow 1, 2, and 3 for β_1 in the simulation.

Table 4.6 shows the selection rates of X_1 variable out of 1,000 iterations for each specified regression model (changing values for β_1). The association between X_1 and X_2 (the correlation coefficients of the X variables are 0.4 for model (M-2a) and 0.25 for model (M-2b)) shrinked the true DIV values towards 0 a little bit, leading to a higher level of instability (compare Table 4.6 with the fourth column of Table 4.2). The fact that X_1 and X_2 in model (M-2a) are correlated more strongly than those in model (M-2b) is reflected in the true DIV's, and in turn in the selection rates. Table 4.6 suggests that, provided that the marginals of X_1 and X_2 are fixed, the higher level of instability is for the larger absolute value of the correlation coefficient.

(Table 4.6 about here)

□

Finally, a simple example follows where we will see how the variation in X can contribute to instability of trees.

Example 4.4

Consider a simple regression model with $r = 1$, and the X variable is binary (0 or 1) with $P(X_1 = 1) = p$. Let the data size be equal to n. Then,

$$\mathbf{X'X} = \begin{pmatrix} n & s \\ s & s \end{pmatrix},$$

where s is the number of the case with $X_1 = 1$ in the data set.

Now,

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} s & -s \\ -s & n \end{pmatrix} (ns - s^2)^{-1}$$

yielding

$$V(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{n} E\left(\frac{1}{\frac{s}{n}\left(1 - \frac{s}{n}\right)}\right). \tag{4.3}$$

By Jensen's inequality,

$$V(\hat{\beta}_1) \geq \frac{\sigma_\epsilon^2}{n}\left(E\left(\frac{s}{n}\left(1 - \frac{s}{n}\right)\right)\right)^{-1} = \frac{\sigma_\epsilon^2}{n} \cdot \left(p(1-p)\left(1 - \frac{1}{n}\right)\right)^{-1} \tag{4.4}$$

From (4.3), we can say that $V(\hat{\beta}_1)$ increases as p approaches 0 or 1 for a given σ_ϵ^2. The inequality in (4.4) provides us with the greatest lower bound of $V(\hat{\beta}_1)$ for the given distribution of the X_1 variable. □

In this section, our purpose was to see some patterns of instability, and we considered some simple regression models. The regression models with larger r would complicate our problem with only a little more gain, since the variable selection is essentially by pairwise comparisons of the IV's.

It is to be noted at this point that the instability discussed in this paper is confined to a node of a tree, not over a whole tree. However, to understand the instability of trees do we need to understand the instability at each node.

In this section, we have seen, for a regression model with r=2,

(1) that the instability level increases as the absolute value of the true DIV decreases,

(2) that the instability due to the noise to the regression model can be cured by increasing the sample size only,

(3) that when the absolute value of the true DIV is small (less than 0.1), increasing the sample size will be of little help; on the other hand, when the absolute value is not less than 0.4, the instability level looks good (the selection rate of X_1 is over 0.8) for the sample size around 30, and

(4) that if we compare the instability levels from any two regression models, both of which are the same except that the X's are independent in one model, and not for the other, then the instability level may be lower for the independent case than for the other case.

Relationship between the DIV or IV and the instability level at each node seems to deserve further study.

5. DISCUSSION.

At the outset, the consideration of the tree approach for a data set obtained from a linear regression model may sound like nonsense. However, if all the regressor variables involved are finitely discrete, then fitting a regression model is equivalent to partitioning the sample space generated by the regressor variables involved in the model fitting. If the same set of regressor variables that are involved in the model fit is used in the tree approach, then the derived tree, in general, gives rise to a partition of the sample space coarser than the one corresponding to the regression approach. This is an advantage of the tree approach over the classical regression approach as far as the prediction accuracies are of an equivalent level.

Many criteria are developed for choosing the best regression models (Seber(1977), Miller (1990)). Among them are the coefficient of determination (R-square), Mallows' C_p, and MSEP. Any of these seems hardly applicable to selection of the final tree. If we have a careful look at the expressions (2.5), (3.6), and (3.7), we can see that the tree is determined by the ls estimate of β, and the relation among the X-variables. In regression,

the estimate of $\underset{\sim}{\beta}$ changes for different sets of regressors; while, in the tree-approach, we use the same estimate of $\underset{\sim}{\beta}$ all through the tree-construction process.

Instability of the tree structure is certainly a drawback in the tree approach, but it also is a signal for further investigation for a sound interpretation of the stochastic properties behind data. Based on the theoretical results and the examples of this paper, I can safely say the followings:

(1) If instability is seen near the bottom of a tree, it may be due to the pure noise in data. Increasing the sample size may help.

(2) If instability is elsewhere, it may be due to association among the regressor or predictor variables. In this case comparing different tree-structures may help for a better insight into the nature behind data.

Instability at a node near the top would affect the whole tree-structure and tree-shape. If the IV's of a set of regressors are more or less at the same level, the instability level may decrease at a very slow rate (see, for example, Table 4.3). Thus even for large sized data, it is not very surprising to see instability. In such a situation, those trees that show up at comparable frequencies (suppose we repeat random subsampling from a data set generated from a statistical model and constructing trees based on the subsampled data lots of times) may deserve equal attention for a sound interpretation of the stochastic properties behind data, since those competing regressor variables may equally be informative for the predicted or dependent variable. In this context, a computer program that can construct a tree where a particular regressor variable is split at a user specified node of the tree is desirable. With

this program, we can construct several trees from a data set, and use them for better interpretation of the stochastic properties behind the data.

References

Atiqullah, M. (1962). The estimation of residual variance in quadratically balanced least squares problems and the robustness of the F-test. *Biometrika*, **49**, 83-91.

Breiman, L, Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth International Group, Belmont, CA.

Doyle, R. M. (1973). The use of automatic interaction detector and similar search procedures. *Operational Res. Quart.*, 24, 465-467.

Einhorn, H. (1972). Alchemy in the behavioral sciences. *Pub. Op. Quart.*, 36, 367-378.

Huang, M. C. (1989). *Piecewise linear tree-structured regression.* Unpublished Ph. D. thesis. Dept. of Statistics, University of Wisconsin-Madison.

Lehmann, E. L. (1983). *Theory of Point Estimation*. John Wiley & Sons, Inc.

Miller, A. J. (1990). *Subset Selection in Regression*. Chapman and Hall, London, England.

Morgan, J. N. and Messenger, R. C. (1973). THAID: A sequential search program for the analysis of nominal scale dependent variables. Ann Arbor: University of Michigan, Institute for Social Research.

Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. J. A. S. A., **58**, 415-434.

Seber, G. A. F. (1977). *Linear Regression Analysis*. John Wiley & Sons, Inc.

Figure 1.1

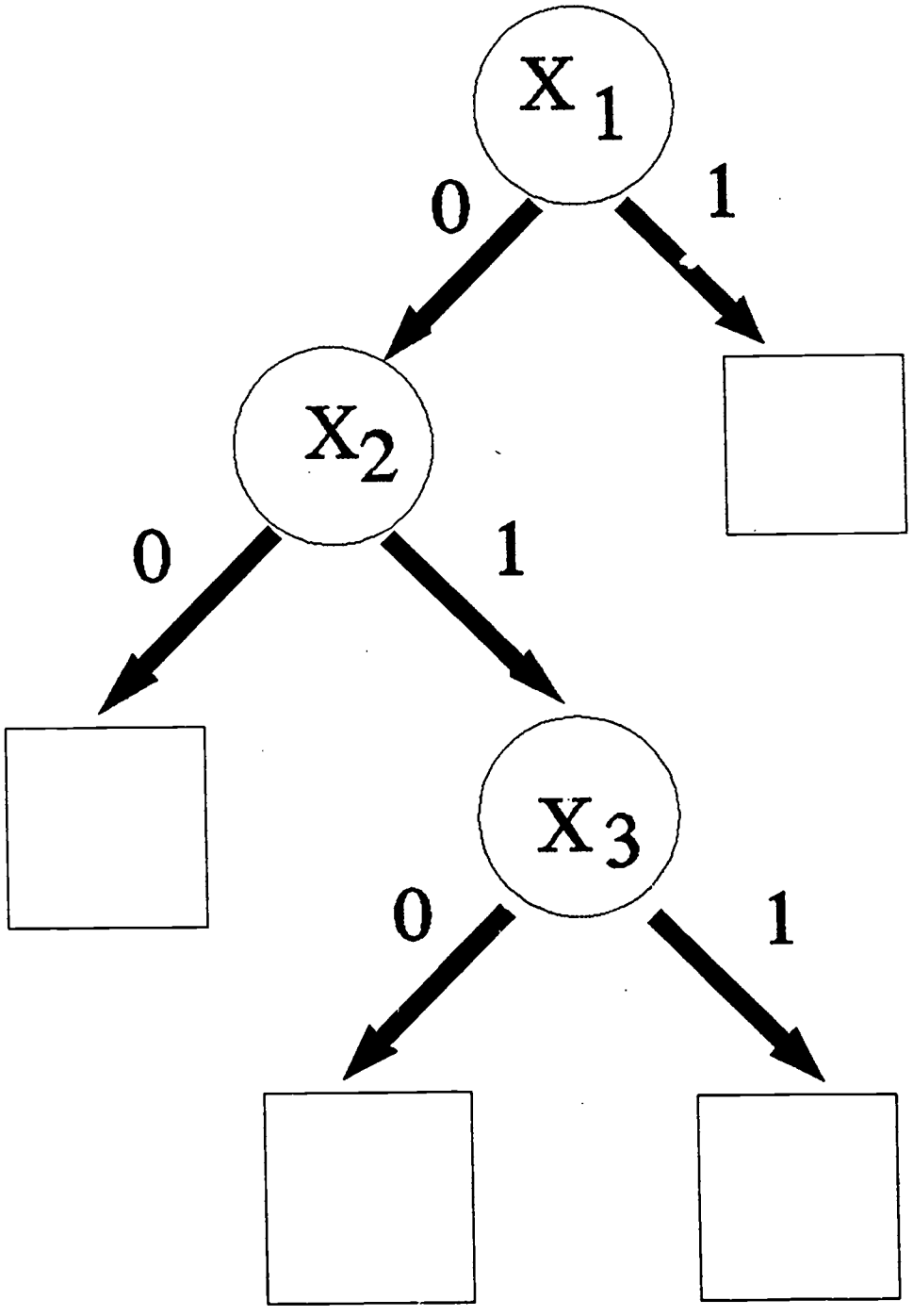

TABLE 4.1

DIV	Variable-selection by CART
0.016	X_1
-0.3	X_2
-0.011	X_2
-0.003	X_2
-0.345	X_2
-0.224	X_2
0.056	X_1
0.033	X_1
-0.047	X_2
-0.218	X_2

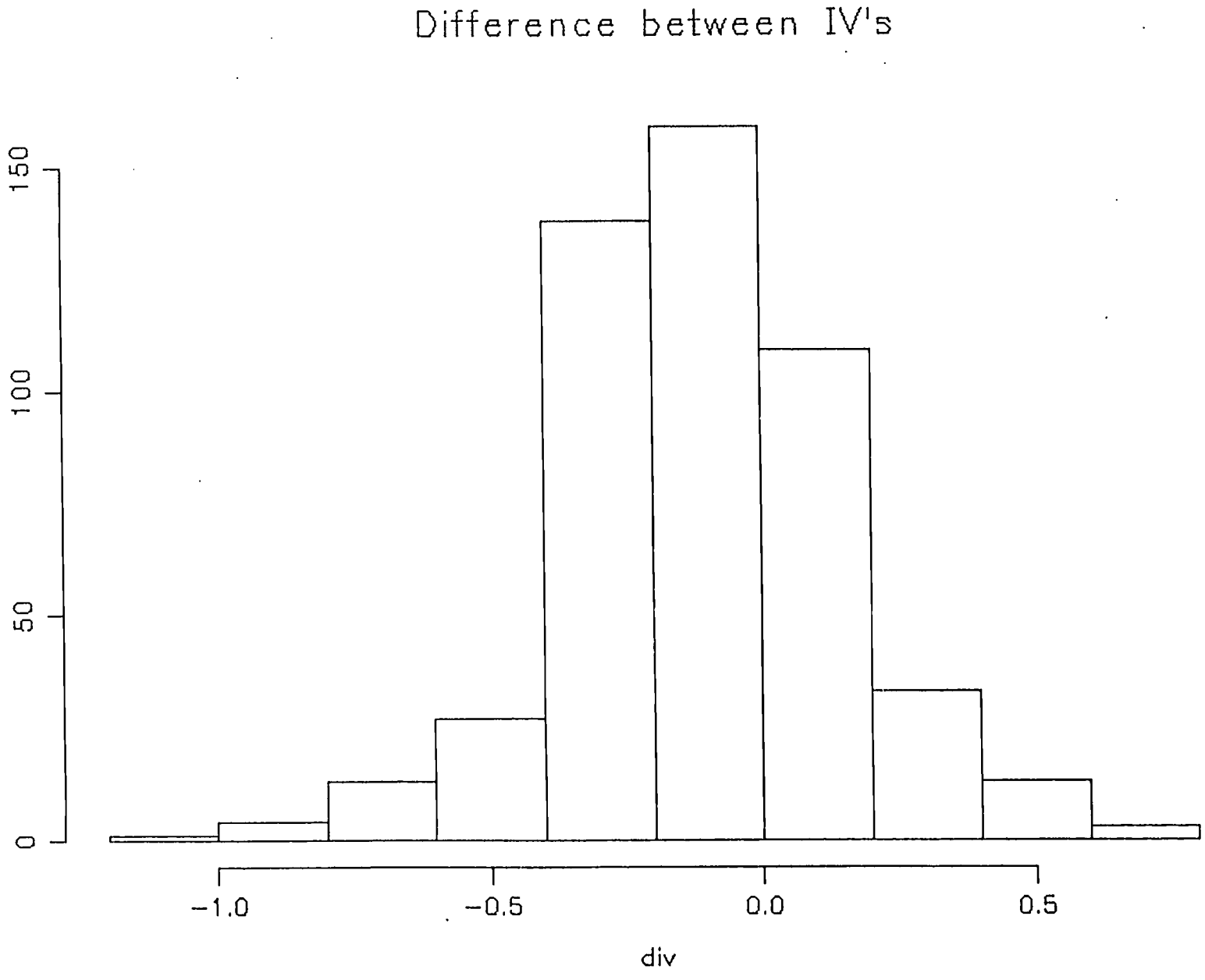

Figure 4.1

Table 4.2

β_1	Sample Size	$(P(X_1 = 1), P(X_2 = 1))$			
		(0.2, 0.3)	(0.2, 0.5)	(0.3, 0.4)	(0.3, 0.5)
1	5	0.52	0.50	0.50	0.50
	10	0.49	0.43	0.46	0.48
	30	0.41	0.37	0.45	0.43
	50	0.38	0.32	0.42	0.44
	E(DIV)	-0.05	-0.09	-0.03	-0.04
2	5	0.74	0.71	0.74	0.74
	10	0.78	0.75	0.80	0.81
	30	0.88	0.85	0.93	0.94
	50	0.93	0.91	0.97	0.97
	E(DIV)	0.43	0.39	0.6	0.59
3	5	0.89	0.90	0.91	0.89
	10	0.93	0.92	0.95	0.96
	30	0.98	0.98	1.00	1.00
	50	1.00	1.00	1.00	1.00
	E(DIV)	1.19	1.15	1.66	1.65

Figure 4.2

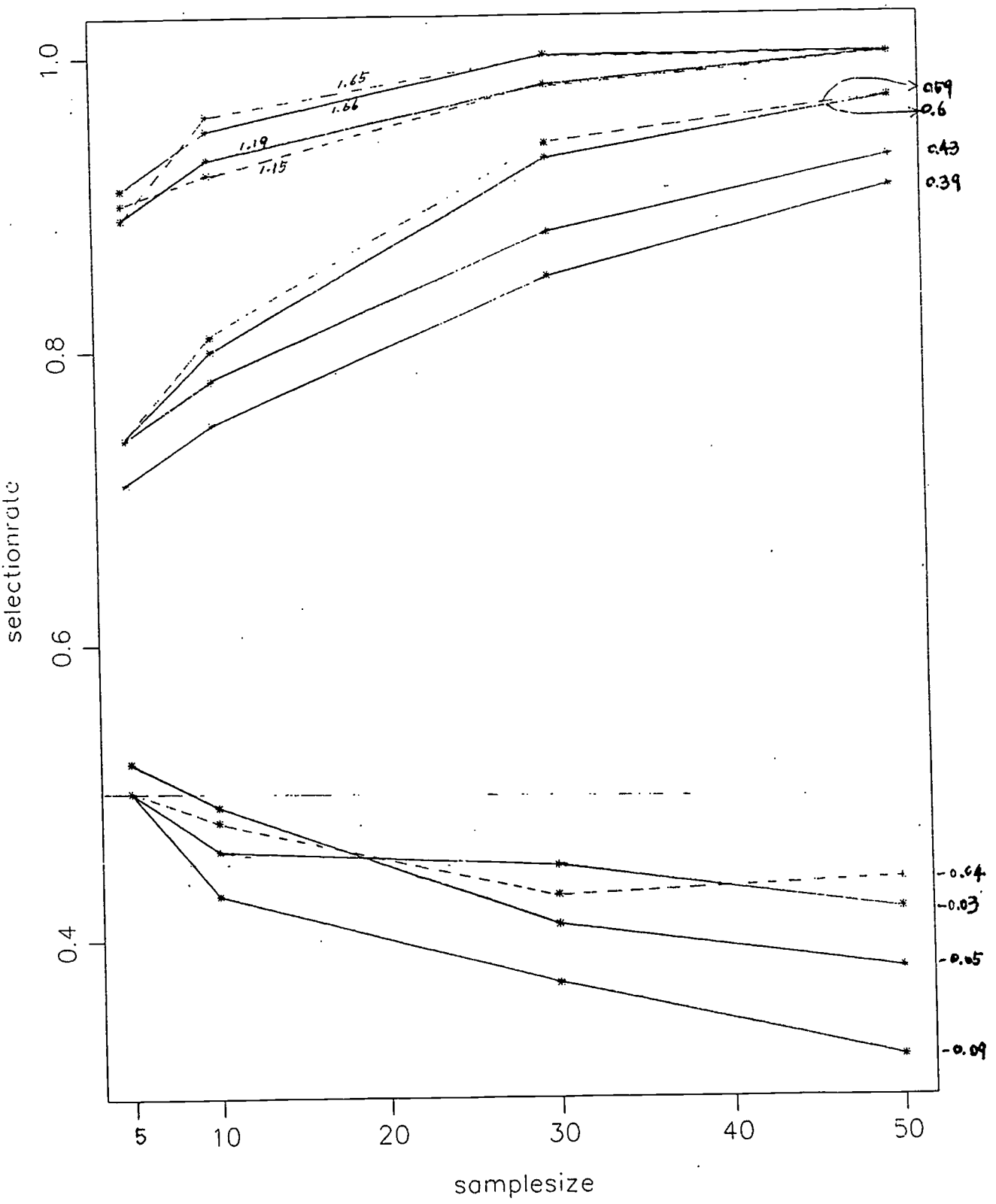
selectionrate
1.0
0.8
0.6
0.4
1.65
1.66
1.19
1.15
0.69
0.6
0.43
0.39
0.04
0.03
0.05
0.09
5
10
20
30
40
50
samplesize

Table 4.3

Sample Size	Selection Rate of X_1
5	0.5
10	0.43
30	0.37
50	0.32
75	0.26
100	0.23
150	0.2
200	0.17
300	0.12
400	0.09
500	0.07
600	0.043
700	0.03
800	0.023
900	0.022

Table 4.4

	Sample Size	σ_ϵ		
		2	3	4
$\beta_1 = 1$	5	0.497	0.49	0.524
	10	0.466	0.469	0.51
	30	0.435	0.438	0.451
	50	0.381	0.456	0.418
$\beta_1 = 2$	5	0.584	0.539	0.548
	10	0.622	0.614	0.542
	30	0.695	0.648	0.646
	50	0.781	0.69	0.636
$\beta_1 = 3$	5	0.716	0.623	0.584
	10	0.764	0.681	0.665
	30	0.887	0.823	0.763
	50	0.952	0.867	0.798

Table 4.5

X_1 \ X_2	0	1
0	0.48	0.32
1	0.02	0.18

(a)

X_1 \ X_2	0	1
0	0.45	0.35
1	0.05	0.15

(b)

Table 4.6

β_1	Sample Size	Selection Rate	
		M-2a	M-2b
1	5	0.48	0.51
	10	0.475	0.47
	30	0.4	0.35
	50	0.36	0.34
	E(DIV)	-0.076	-0.084
2	5	0.66	0.7
	10	0.695	0.697
	30	0.79	0.805
	50	0.87	0.89
	E(DIV)	0.33	0.366
3	5	0.82	0.87
	10	0.87	0.89
	30	0.97	0.975
	50	0.99	0.996
	E(DIV)	1	1.12